



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-442951

Analysis of the Noise in Data from the Mt. Meron Array

D. H. Chambers, E. Breitfeller

July 16, 2010

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Analysis of the Noise in Data from the Mt. Meron Array

David H. Chambers, Eric Breidfeller

Introduction

This memo describes an analysis of the noise in data obtained from the Mt. Meron seismic array in northern Israel. The overall objective is to develop a method for removing noise from extraneous sources in the environment, increasing the sensitivity to seismic signals from far events. For this initial work, we concentrated on understanding the propagation characteristics of the noise in the frequency band from 0.1 – 8 Hz, and testing a model-based method for removing narrow band (single frequency) noise.

The Mt. Meron seismic array is an array of 16 sensors covering a geographical area of around 6 square km. and an elevation range of 200 m in northern Israel. Figure 1 shows a plan view of the array and a perspective view obtained from GoogleEarth. The array elements are arranged roughly on a tilted plane whose elevation varies from 701 meters at element MMC2 to 893 meters at element MMC6. Each sensor measures the vertical displacement at a sampling rate of 40 Hz (see ref. 1 for more detail). Two data sets consisting of 9 and 10 contiguous days were analyzed. The first set was measurements taken Jan 1-9, 2008 (9 days) while the second set was taken Mar 20-29, 2008 (10 days). Both broadband and narrowband (single frequency) noise sources were observed. The narrowband noise sources presumably come from rotating or reciprocating machinery close to the array. The broadband noise sources may be either man-made (*e.g.* plant noise), or natural sources (ocean generated wave noise).

Two types of analysis were performed on the data. The goal of the first type of analysis was to characterize the noise and seismic environment. Spectrograms and cross-spectra were calculated to identify noise components and determine wave speed and direction. The second type of analysis built and applied a time-domain beamforming algorithm for removing narrowband noise components. Unlike more conventional beamforming algorithms, this new algorithm attempts to track the source location and frequency of the noise using a model of the noise source. The algorithm is a generalization of the classical Kalman filter that can accommodate nonlinear signal models and non-Gaussian statistics. Once the algorithm determines the best fit to the model, it uses the model to predict the noise signal and remove it from the data.

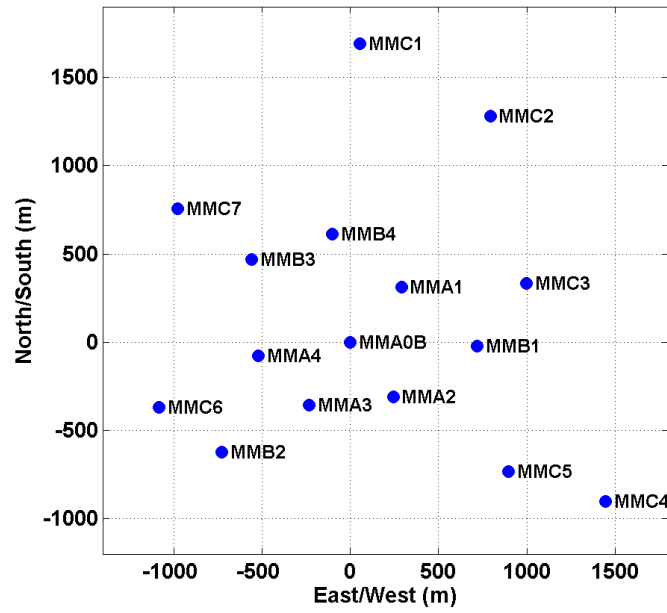


Figure 1

Layout (top) and GoogleEarth perspective view (bottom) of Mt. Meron Array.

Noise Characterization

The goal of the first type of analysis performed on the Mt. Meron data was to identify the noise components and determine the speed and direction of the seismic waves in that environment. Spectrograms of the data from each array element were calculated for both data sets. Figure 3 shows typical spectrograms from the Jan2008 and Mar2008 data sets. These were calculated using Welch's method to estimate the power spectrum over 10 minute blocks of data (nonoverlapping). There are several components to the background noise visible in the spectrograms. One component is a broadband background noise that varies on a diurnal cycle. This is characteristic of traffic and industrial noise associated with the local economy. In fact the third and tenth days in the March data and the fifth day in the January data shows a noticeable drop in background noise associated with observing the Sabbath. The second most noticeable component is a low frequency (< 1 Hz) band of noise that is persistent in the Jan2008 data set but arises only on the sixth day in the Mar2008 data. This noise overlaps the frequency range where long-range seismic signals are expected, limiting the overall performance of the array. The third noise component is a series of narrowband sources likely associated with rotating machinery such as pumps. Some of these frequency lines are intermittent while others show frequency wandering. These lines occur in the higher frequency range where short-range seismic signals might be observed. They could be removed using a model-based processing scheme designed to track the frequency of each source and estimate its location. This is described in the second section of this report. An examination of the variation of the noise with sensor location (see appendix) shows that most of the noise components vary little across the array. An exception is the intermittent high frequency (8-9 Hz) noise, which is noticeably greater for sensors MMC1 and MMC7 close to the community of Sasa.

In order to apply beamforming algorithms for noise removal, we need to know the wave speed for each range of frequencies. Our approach was to estimate the phase difference between each pair of elements from the cross-spectrum (Fourier transform of the cross-covariance matrix). The cross-spectrum was estimated using a variation of Welch's method for power spectra. Products of the Fourier transformed signals from each sensor were averaged over 20 minute time blocks, giving cross-spectrum estimates at 20 minute intervals over each data set. Mathematically, the cross-spectrum between sensors m and n is $C_{mn}(\omega) = E\{\hat{u}_m(\omega)\hat{u}_n^*(\omega)\}$, where $\hat{u}_m(\omega)$ is the Fourier transform of the signal from sensor m and $E\{\cdot\}$ is the expectation. Each cross-spectrum covered the frequency range of 0 to 8 Hz at a resolution of around 0.02 Hz. This was used to analyze the noise characteristics in the frequency band above 1 Hz. For frequencies less than 1 Hz, a low range version of the cross-spectrum was calculated for 0-1 Hz with a resolution of about 0.002 Hz for each 20 minute block. Formally, the cross-spectrum for a 16 element array has 256 components (16 x 16). However, symmetry reduces these to 136 independent components, 16 of which are the ordinary power spectra for each element.

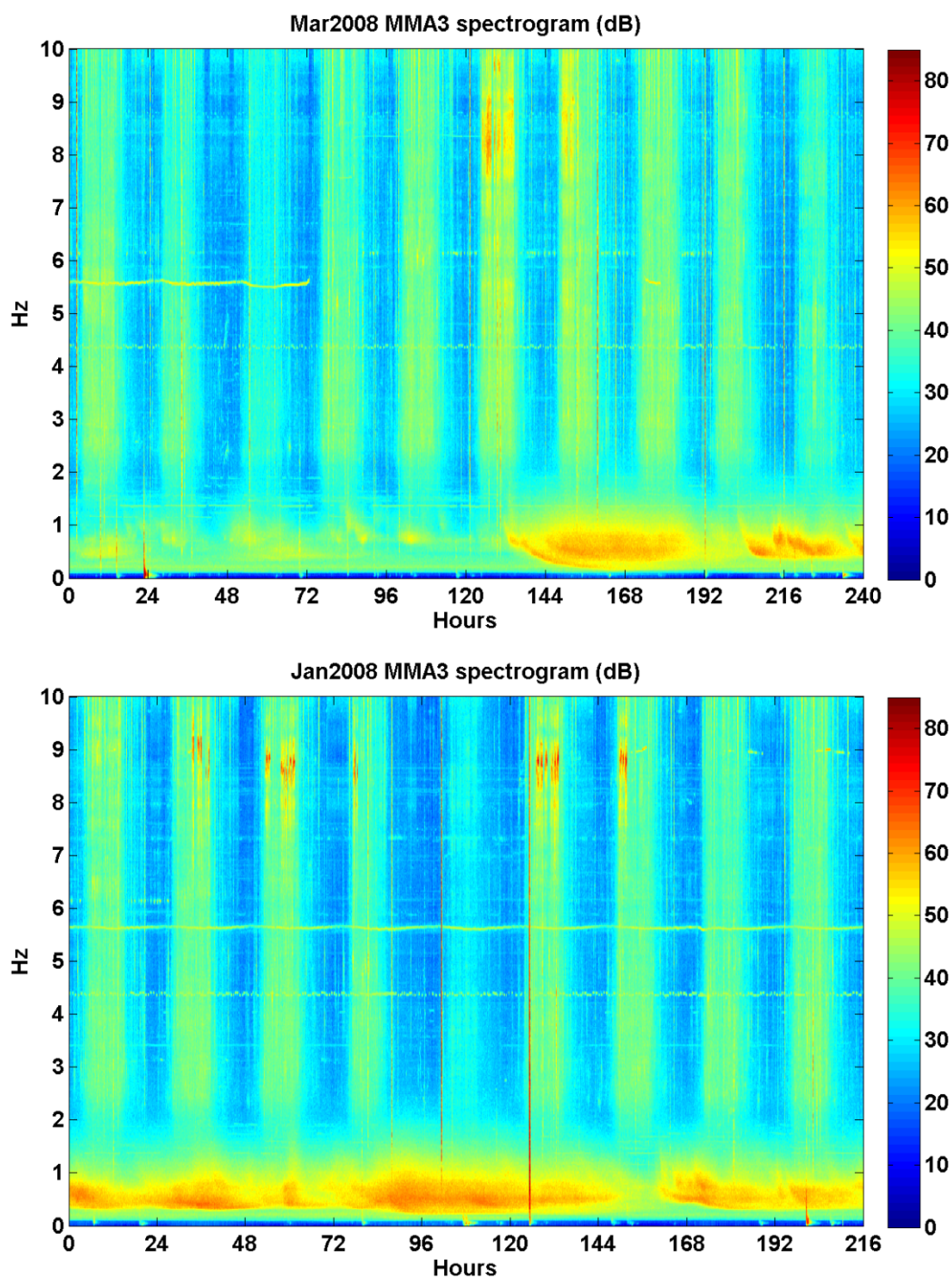


Figure 2
Spectrogram of array element MMA3 for March 20-29, 2008 (top) and January 1-9, 2008 (bottom).

The components of the cross-spectrum can be considered spatial samples of a continuous cross-spectral density function, $C_{mn}(\omega) = C(\Delta x_{mn}, \Delta y_{mn}, \omega)$, where $\Delta x_{mn} = x_m - x_n$, and $\Delta y_{mn} = y_m - y_n$. A plot of the spatial separations (Fig. 3) shows that the minimum correlation distance that can be calculated is around 500 m, and the maximum is nearly 2500 m. The directional distribution is fairly uniform, detecting correlations in any direction. This is important in estimating the direction of propagation of noise components.

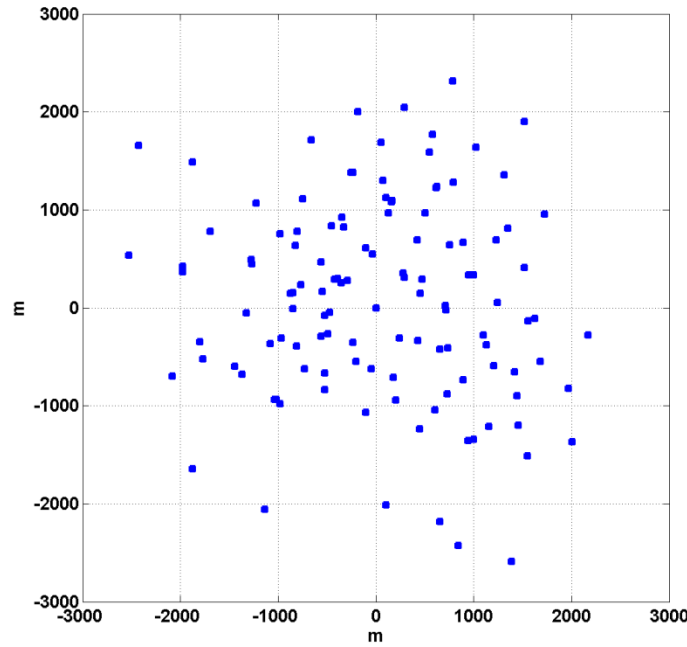


Figure 3

Spatially sampling of the cross-spectral density. The horizontal axis is Δx_{mn} and the vertical axis is Δy_{mn} .

Suppose the noise consisted of a random superposition of plane waves at different frequencies ω . Then the phase of the cross-spectrum $C_{mn}(\omega)$ would be $\Phi_{mn}(\omega) = \alpha_{mn} + k\Delta x_{mn} \cos(\theta_k) + k\Delta y_{mn} \sin(\theta_k)$, where $k = \omega/c = 2\pi f/c$, α_{mn} is a constant, and θ_k is the propagation direction of the wave component at frequency ω . For a fixed frequency this forms a plane in the space of phase Φ versus Δx and Δy . To test this we reduced the number of cross-spectra in each data set by averaging across multiple days to produce a sequence of cross-spectra over a single day period. For the Jan2008 data this reduced the number of cross-spectra from 648 (one for each 20 minute period over 9 days) to 72. Each of the 72 cross-spectra is an average of nine cross-spectra that are chosen at the same time in the diurnal cycle. This also reduces the effect of transients, leaving the persistent noise sources in each data

set. These were further averaged to a single cross-spectrum for display purposes. Figure 4 shows the cross-spectrum phase for three different frequencies below 1 Hz and a least-squares plane fit for each. The phases appear to be distributed on parallel planes, with most of the phases on the central plane. The upper and lower planes correspond to phases $\pm\pi/2$ from the central plane. In Fig. 4, the phases for the highest frequency (0.75 Hz) are less concentrated around planes than the lower frequencies. This is a general property of the higher frequencies, leading to greater uncertainty in the properties of the noise at higher frequencies.

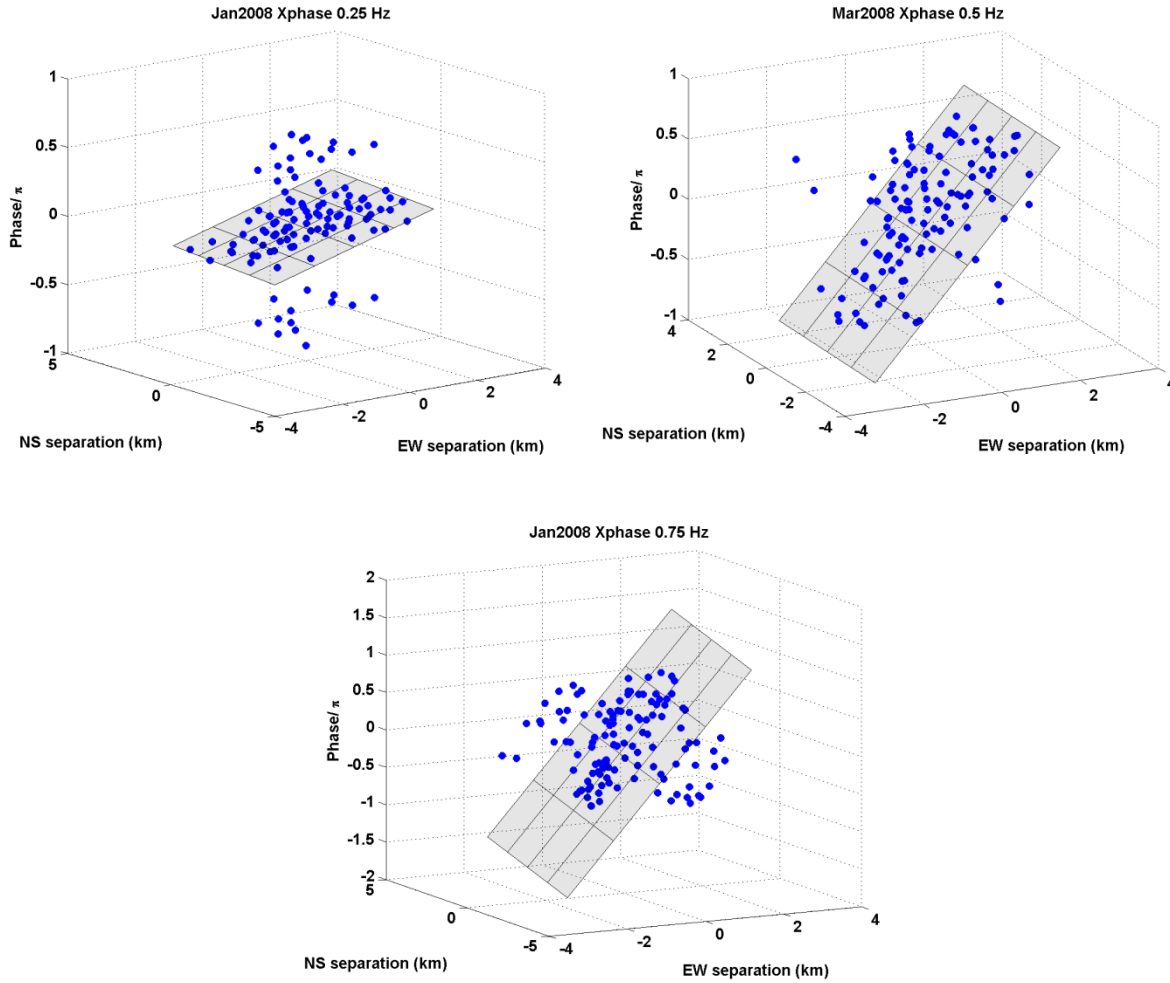


Figure 4

Cross-spectrum phase versus separation for three frequencies: 0.25 Hz (upper left), 0.5 Hz (upper right), and 0.75 Hz (bottom). These results are typical for both the Jan2008 and Mar2008 data sets.

From the coefficients of the least-squares planar fit, we can estimate the propagation direction of the noise component and the wave speed at each frequency. Figure 5 shows the wave speed and direction of propagation estimated from the phase of the overall mean cross-spectrum for each data set. Both wave speed and propagation direction vary smoothly and agree well between both data sets for the frequency range of 0.25 – 1.0 Hz. The wave speed ranges from 2100 m/s near 0.65 Hz to around 7000 m/s near 0.25 Hz and 1.0 Hz. The direction of propagation is due east. This would be consistent with noise produced by waves on the Mediterranean shoreline. For frequencies below 0.25 Hz, the wave speed and direction estimates vary rapidly. In particular, the wave speed estimate shows sharp peaks with unphysical values well above 10,000 m/s. Two effects may contribute to inaccurate and rapidly varying estimates in this range. First, the wavelengths at these low frequencies are longer than the aperture of the array, making it harder to obtain accurate phase differences. Second, large apparent wave speeds would be expected if the wave direction has a substantial vertical component. In fact, waves propagating from directly below the array would have infinite apparent wave speeds since the wavefronts would be horizontal. Figure 6 shows the phase of the cross-spectrum at 0.2 Hz for the Jan2008 data set. The least-squares planar fit is nearly horizontal, showing little phase variation over the range of separations. Thus the assumption of a two-dimensional wave field for the noise components may be breaking down at frequencies below 0.25 Hz.

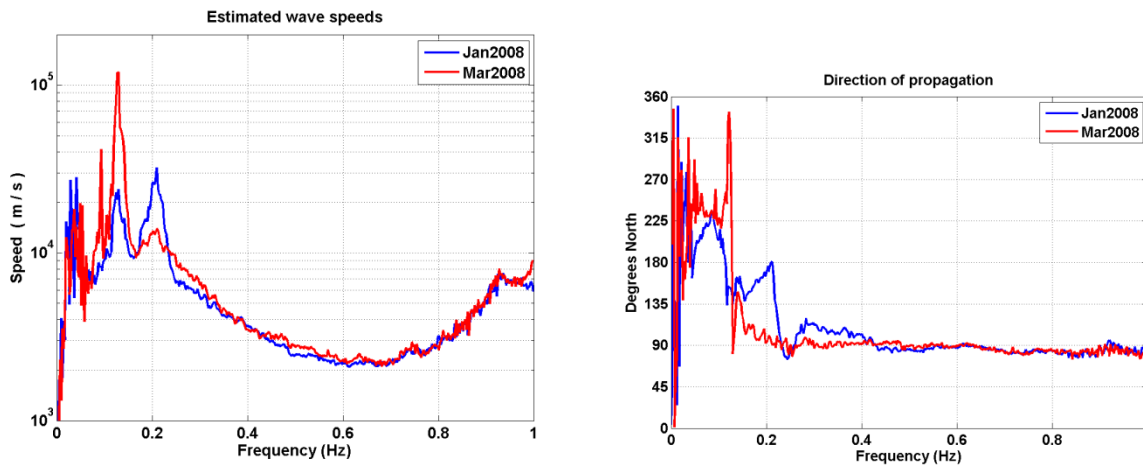


Figure 5

Estimated wave speeds (left) and propagation directions (right) derived from the phase of the mean cross-spectrum for each data set.

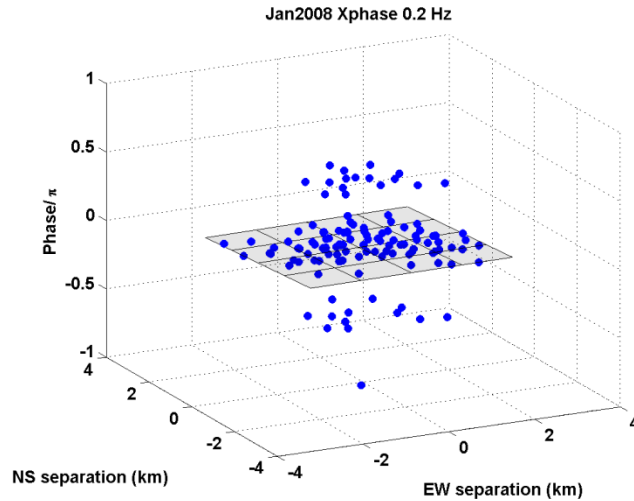


Figure 6

Cross-spectrum phase versus separation at 0.2 Hz for the Jan2008 data set. Note the small variation of phase over the range of separations.

The wave speeds and propagation directions shown in Fig. 5 are averaged over the entire data sets. Figures 7 and 8 show these estimated for each cross-spectrum in the data over several days. For frequencies in the upper range (>0.3 Hz) the estimated wave speed and propagation direction is close to that for the overall average in Fig. 5. The wave speed varies smoothly in a range between 2100 m/s to around 7000 m/s, and the propagation direction is nearly due east. As noted before, there is considerable uncertainty in the estimates for frequencies less than 0.25 Hz due to long wavelengths and possible vertical propagation. The boundary between the higher frequency stable region and the lower frequency high-variability region seems to change over time. This could be caused by a change in the bandwidth of a vertically propagating noise component. It may be possible to characterize this with further work on more data.

So far we have concentrated on noise components with frequencies less than 1 Hz. For higher frequencies, the cross-spectrum approach begins to break down because the wavelengths are less than the 500 m minimum separation distance between sensors. The phase variation between neighboring sensors becomes large and could even be more than one full cycle. For higher frequencies, we concentrate on removing narrowband noise components that are associated with machinery.

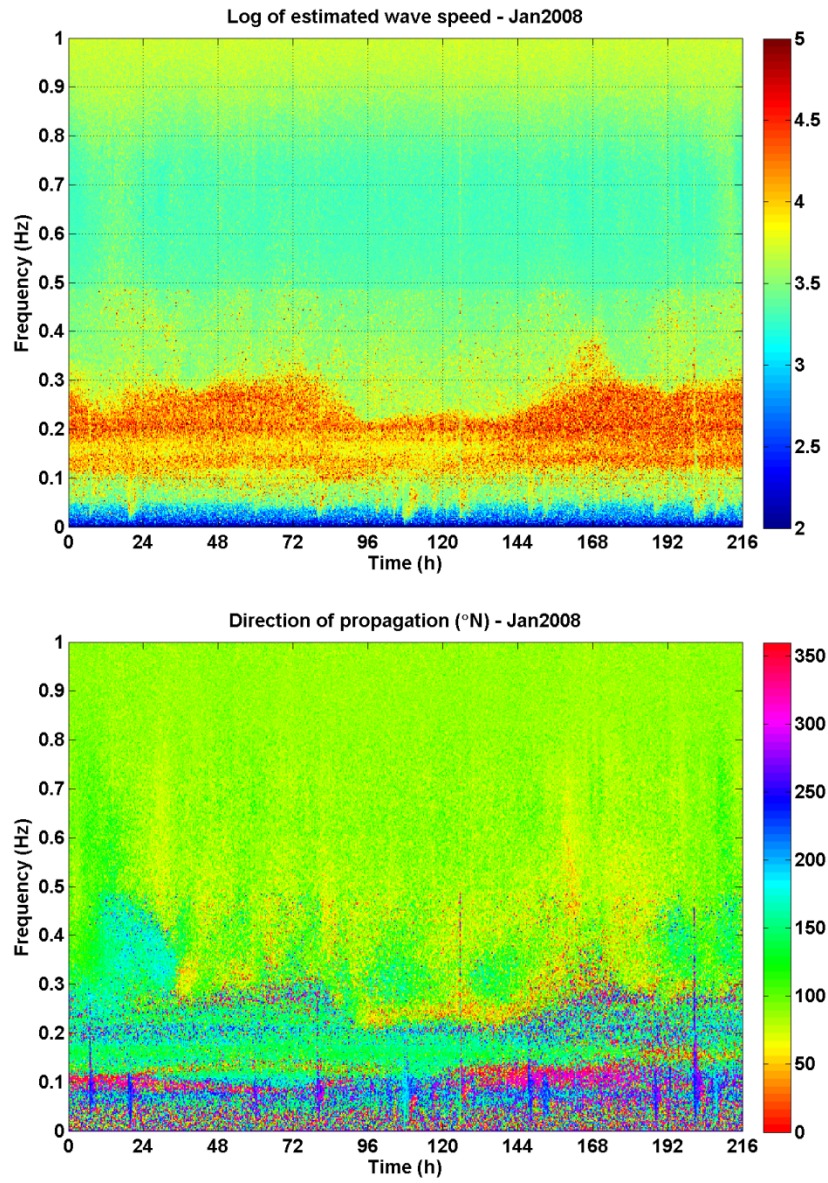


Figure 7

Estimated wave speed (top) and propagation direction (bottom) obtained from the cross-spectrum for the Jan2008 data set. Time sample interval is 20 minutes.

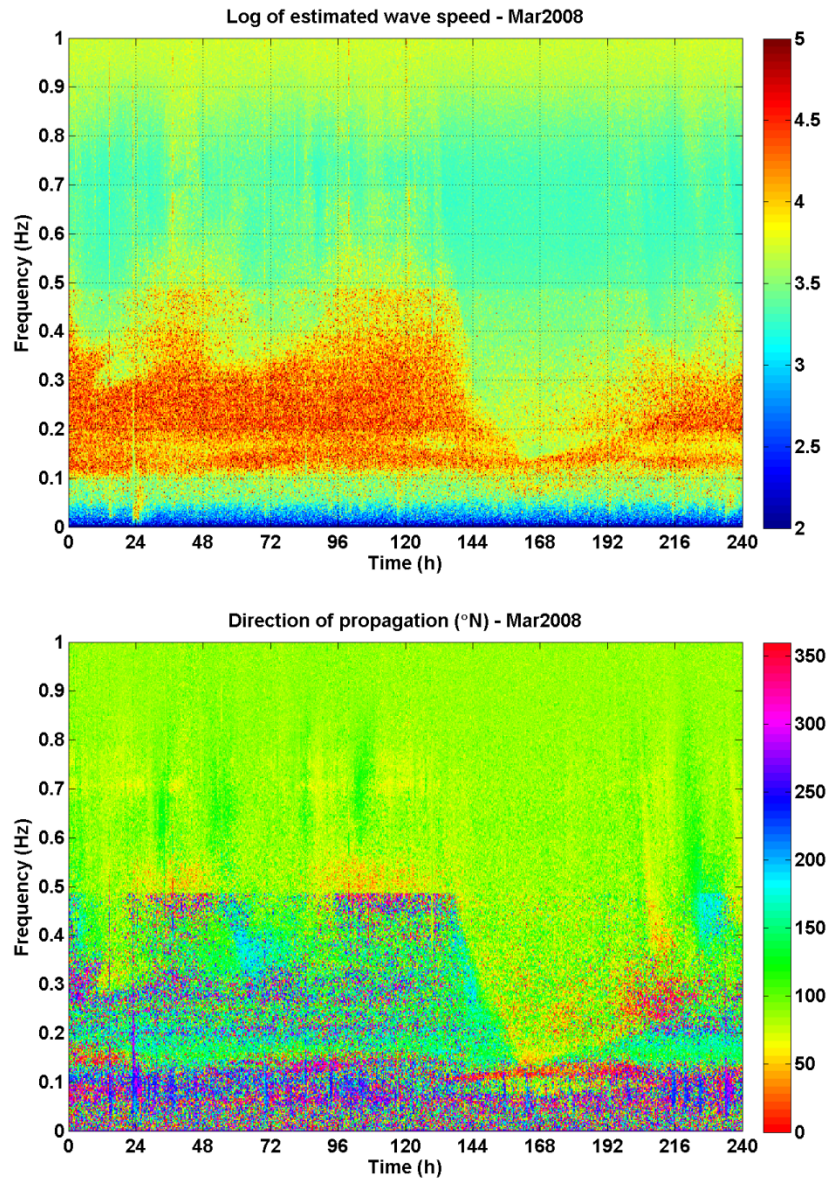


Figure 8

Estimated wave speed (top) and propagation direction (bottom) obtained from the cross-spectrum for the Jan2008 data set. Time sample interval is 20 minutes.

Model-based Noise Identification and Removal

As noted before, one noise component seen in both the Jan2008 and Mar2008 data sets is narrow band tones most likely generated by machinery. The sources for this noise may be far from the array or within the footprint of the array itself. One approach for removing this type of noise, especially if the noise varies in frequency and amplitude, is to fit a mathematical model to the noise. Once the best fit is found, the model is used to subtract out the noise. The particular approach we used, the Unscented Kalman Filter (UKF), is a member of the class of sequential Bayesian methods, which continuously updates a statistical model of the noise as each measurement is made. It is a generalization of the classical Kalman filter that can incorporate nonlinear models and non-Gaussian statistics.

The UKF begins with a mathematical model of the narrow-band noise component. We assume that the noise can be described as a single frequency tone which may be slowly modulated in either amplitude or frequency (or both). Assuming two-dimensional geometry, the source tone is described by an amplitude, frequency, and source location (x and y coordinates). We start with the far-field representation of the field radiated from a point source at (x_s, y_s) in two-dimensions:

$$\psi(x, y, t) = \frac{A}{\sqrt{R(x, y)}} e^{i(kR(x, y) - \omega t)}$$

where $R(x, y) = \sqrt{(x - x_s)^2 + (y - y_s)^2}$ and A is the (complex) amplitude. For convenience choose a reference point (x_0, y_0) near the center of the array, then writing

$R(x, y) = R(x_0, y_0) + \Delta R(x, y)$ the field becomes

$$\psi(x, y, t) = \frac{B}{\sqrt{1 + \Delta R(x, y)/R(x_0, y_0)}} e^{i(k\Delta R(x, y) - \omega t)}$$

with B the new complex amplitude proportional to the original amplitude A . Using the definition of ΔR for computations can be numerically inaccurate. A better expression is

$$\Delta R(x, y) = \frac{(x - x_0)(x + x_0 - 2x_s) + (y - y_0)(y + y_0 - 2y_s)}{R(x, y) + R(x_0, y_0)}.$$

This has the advantage of maintaining accuracy even when the source point is far from the array. Finally, the field is sampled by each sensor in the array (location (x_n, y_n)) so we can write

$\psi_n(t) = \psi(x_n, y_n, t)$, and similarly for R_n and ΔR_n .

To complete the formulation of the model for the UKF we must transform the expression for the field into a state-space model that describes the evolution of the model in time. Since the

model must describe a process that is inherently random, it takes the form of a system of first-order ordinary differential equations with random forcing (Gauss-Markov). The dependent variables, or states, will include the fundamental quantities that describe a source of sinusoidal waves, namely the frequency f , amplitude A , and the source location (x_s, y_s) . Though nominally constant, the model allows for slow time variations for these quantities. This requires adding another state variable, the total phase $\Theta(t)$ whose time derivative is the frequency. The equations for the evolution of the five states are then

$$\frac{d\Theta}{dt} = f = u_1(t), \quad \frac{df}{dt} = u_2(t), \quad \frac{dA}{dt} = u_3(t), \quad \frac{dx_s}{dt} = u_4(t), \quad \frac{dy_s}{dt} = u_5(t),$$

where $u_m(t)$ are white Gaussian noise. To complete the model, we must relate the states to the measurements. We use the expression for the field from a single point source derived above to write the following measurement equation:

$$\psi_n(t) = \frac{A(t)}{\sqrt{1 + \Delta R_n(x_s, y_s)/R_0(x_s, y_s)}} \cos\left(\frac{2\pi f(t)}{c} \Delta R_n - \Theta(t)\right) + v_n(t),$$

where $R_0(x_s, y_s) = \sqrt{(x_s - x_0)^2 + (y_s - y_0)^2}$ and $v(t)$ is white noise. The basic signal processing problem is to estimate the states at each time sample from the sampled measurements. Besides the state variables, the other unknowns that serve as a type of adjustment for the processor are the covariances of the additive noise terms, both in the state equations and the measurement equation.

As an introduction to the Unscented Kalman Filter (UKF) we will first describe the operation of the classical Kalman Filter (Fig.9). The Kalman filter is valid whenever the statistics are Gaussian and the model equations (both state and measurement) are linear in the states. This guarantees that probability distribution for the states remains Gaussian for all times and it is sufficient to estimate the state mean and covariance to describe all the statistical properties of the model. The Kalman filter is sequential estimator, *i.e.* given the means and covariances of the states at time $t-1$ (discrete) it allows one to calculate the means and covariances of the states at time t using the data $\mathbf{y}(t)$. It begins with the vector of state means $\mathbf{x}_{\text{est}}(t-1|t-1)$ and covariance matrix $\mathbf{R}_{\text{est}}(t-1|t-1)$ at discrete time index $t-1$. Equations derived from the state space model are used to predict the state vector mean and covariance matrix at time index t . The measurement equation is then used to predict the measurement. The difference between the predicted measurement $\mathbf{y}_{\text{est}}(t|t-1)$ and the data $\mathbf{y}(t)$ is the innovation $\mathbf{e}(t)$. This, along with the assumed covariances of the noise terms (\mathbf{R}_{uu} and \mathbf{R}_{vv}) in the state and measurement equations, are used to calculate the Kalman gain $\mathbf{K}(t)$. Finally, the Kalman gain is used to correct the predicted state vector mean and covariance matrix. This completes the cycle for a single time step. The process is repeated for subsequent time steps. To apply the Kalman filter to a set of data requires choosing initial values for both the state variables and their covariance. The filter is “tuned” by

the choice of the noise covariance matrices \mathbf{R}_{uu} and \mathbf{R}_{vv} . Typically, the initial values are chosen from an initial assessment of the data or from physical arguments. The noise covariances are the primary “knob” for controlling the performance of the filter. Often a preliminary analysis of the data provides guidance for the choice of measurement noise covariance \mathbf{R}_{vv} . This leaves the state noise covariance \mathbf{R}_{uu} as the primary model parameter that is adjusted to optimize the performance of the filter.

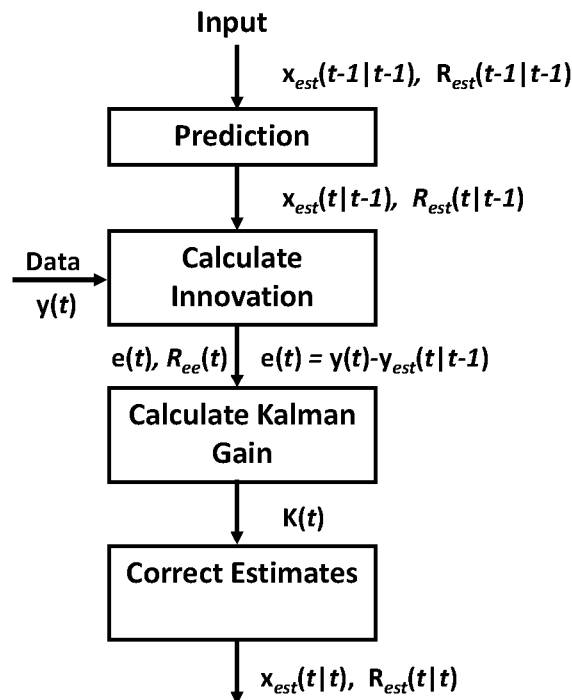


Figure 9

Kalman filter algorithm for updating the state mean and covariances.

The classical Kalman filter requires linear state and measurement equations and Gaussian statistics. Though the model equations for describing a slowly varying sinusoid are linear, the measurement equation is nonlinear. Consequently, even if the states are statistically Gaussian, the measurement is unlikely to be Gaussian. The UKF is designed to overcome these obstacles, allowing states and measurements to be statistically non-Gaussian within certain general limitations, and enabling nonlinear state and measurement equations. The key is a discrete representation of the probability density function (pdf) of the states and measurements that is designed to be exact for calculating means and covariances, but only approximate for higher order moments. The representation for a particular multivariate pdf $p(\mathbf{x})$ is given by a set of *sigma* points \mathbf{x}_σ and corresponding weights \mathbf{W}_σ . These are also used to derive a sigma point representation for the pdf of $\mathbf{y} = f(\mathbf{x})$, in which $\mathbf{y}_\sigma = f(\mathbf{x}_\sigma)$. The sigma points and weights are

chosen so that combinations of them give the means and covariances for both \mathbf{x} and \mathbf{y} . Expressions for higher order moments in terms of the sigma points and weights are only approximate. The representation is restricted to unimodal distributions such as Gaussian, Rayleigh, and Poisson. For more details see the book by Candy [1]. For the particular state and measurement equations we use to model a modulated sine wave, the states are Gaussian random variables since the noise terms $u_m(t)$ are white Gaussian noise. However, the measurements $\psi_n(t)$ are non-Gaussian because they are related nonlinearly to the states. The UKF algorithm proceeds like the classical Kalman filter except the state variables and covariances are replaced by the sigma point values and weights, which are updated at each iteration. Whenever the actual state variables or covariances are required, they are calculated from the sigma points and weights.

We first applied the UKF to simulated data using the Mt Meron array layout. To simplify the system, we set the source amplitude to unity and added white Gaussian noise at a set dB level to create the data. We can then eliminate the amplitude state variable $A(t)$ and reduce the measurement equation to

$$\psi_n(t) = \frac{1}{\sqrt{1 + \Delta R_n(x_s, y_s)/R_0(x_s, y_s)}} \cos\left(\frac{2\pi f(t)}{c} \Delta R_n - \Theta(t)\right) + v_n(t).$$

We also let the wave speed c be a state variable, allowing the processor to estimate the best value from the data. The final set of state equations for the simulation were

$$\frac{d\Theta}{dt} = f = u_1(t), \quad \frac{df}{dt} = u_2(t), \quad \frac{dc}{dt} = u_3(t), \quad \frac{dx_s}{dt} = u_4(t), \quad \frac{dy_s}{dt} = u_5(t).$$

Figure 10 shows the array reference point (green dot) and the source location (red dot) for a particular simulation with source frequency of 0.5 Hz. The reference location is the numerical average of the array sensor coordinates. The data sampling rate was 40 Hz and the wave speed was 2300 m/s. To show how the UKF locks on to the source frequency and location we start the sinusoid 383 seconds after the beginning of the data set. Figure 11 shows a typical simulated sensor channel for a signal-to-noise ratio (SNR) of 10 dB. The performance of the UKF processor for this case is shown in Figure 12. During the time before 383 seconds where the input is white noise, all the states except the frequency remain near their initial values. The frequency rapidly drifts to zero during this period. After the sinusoidal signal begins, the frequency approaches the true value in around 100 seconds. The wave speed and source location states converge more slowly, taking nearly 20 minutes to approach the true values. Figure 13 shows the time history of the square root of the estimated variance for each state. Before 383 seconds, the UKF is reinitialized whenever the estimated variance exceeds a set threshold. This prevents the state values from wandering outside a physically meaningful range when no signal is present.

The UKF was tested on simulated data for sources in a number of locations. The results for a source at position (1230, 2060) meters is shown in Fig. 15. We also delayed the start of the signal to 766 seconds after the beginning of the data record. The UKF position estimates converged to the correct positions in about 40 minutes.

Though the UKF successfully converged using simulated data, it did not converge to a steady solution on real data in the frequency range from 2 to 8 Hz. Though there was insufficient time to do a thorough analysis, there are at least two possible reasons for the failure of the UKF to converge. One may be that the data may need preconditioning before applying the UKF in order to narrow the frequency range of interest. This may be accomplished by applying a bandpass filter around the noise frequency. Another reason may be that the waves in the frequency range of interest may be under-resolved, *i.e.* the sensors may not be close enough together to resolve the wave field for these frequencies. In that case there are multiple possible source locations and the UKF would not be able to decide between each possible solution.

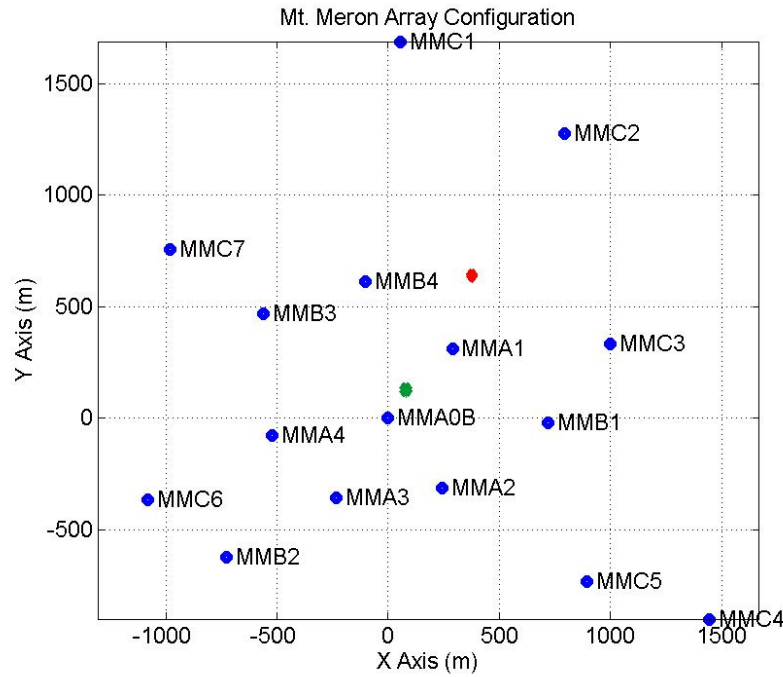


Figure 10

Sensor locations for Mt Meron array with reference location $(X, Y) = (80, 130) \text{ m}$ (green dot) and simulated source location $(X, Y) = (380, 640) \text{ m}$ (red dot).

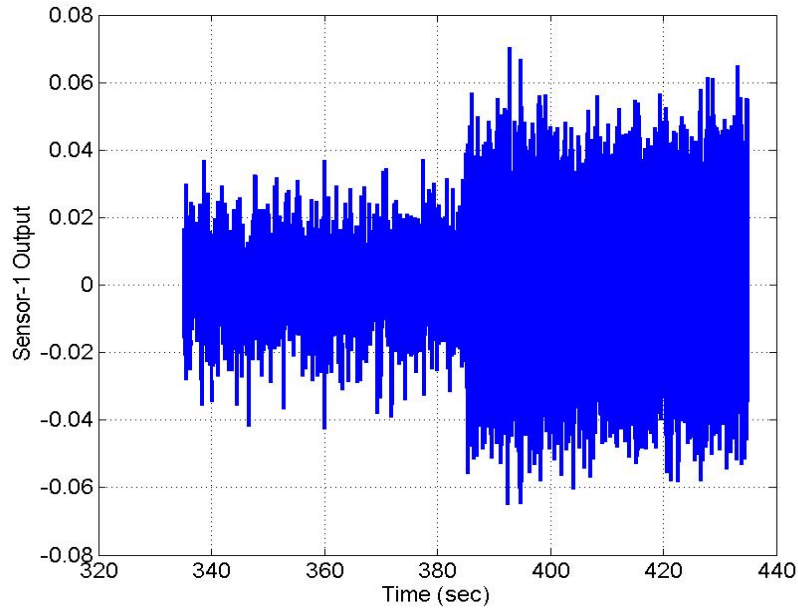


Figure 11

Simulated data for a single sensor at an SNR of 10 dB.
The sinusoidal signal begins at 383 sec.

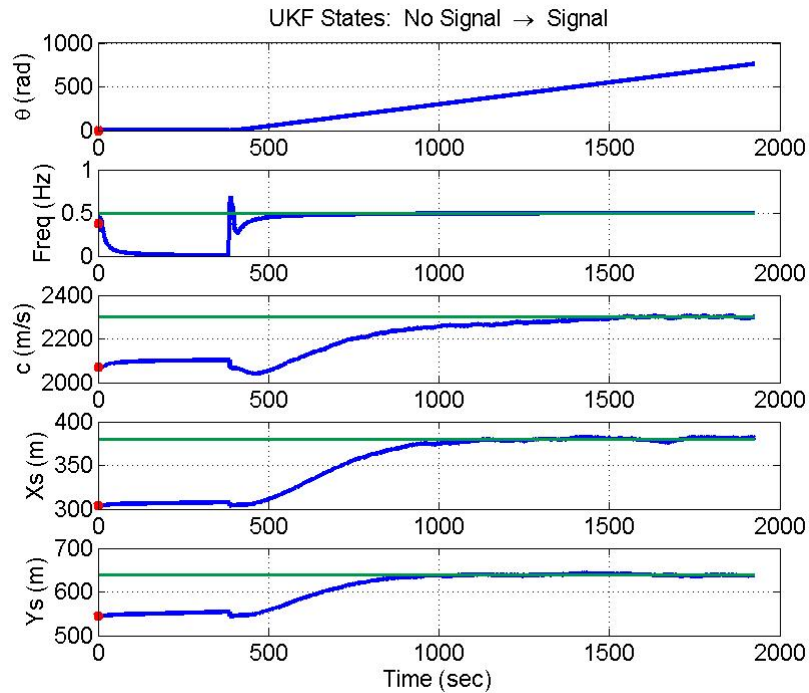


Figure 12

Time history of the states when the signal is turned on 383 seconds after the beginning of the record. The red dots at time zero are the initial state values. The green reference lines are the true values for each state.

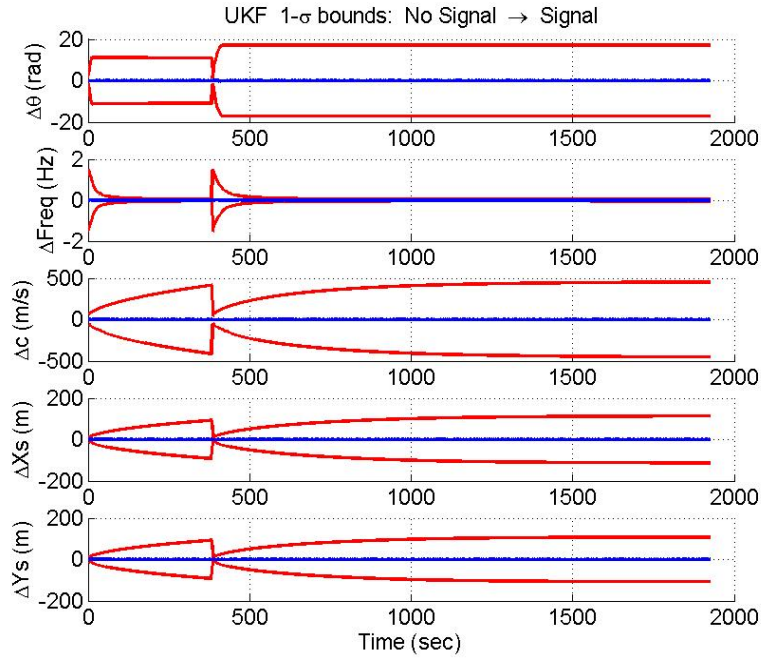


Figure 13

Time history of the residuals (innovations) for each state (blue) and $\pm\sigma$ (red). For times before the signal is present (before 383 seconds) the UKF is re-initialized whenever the estimated variance exceeds a threshold. This keeps the states from wandering too far when there is not signal present.

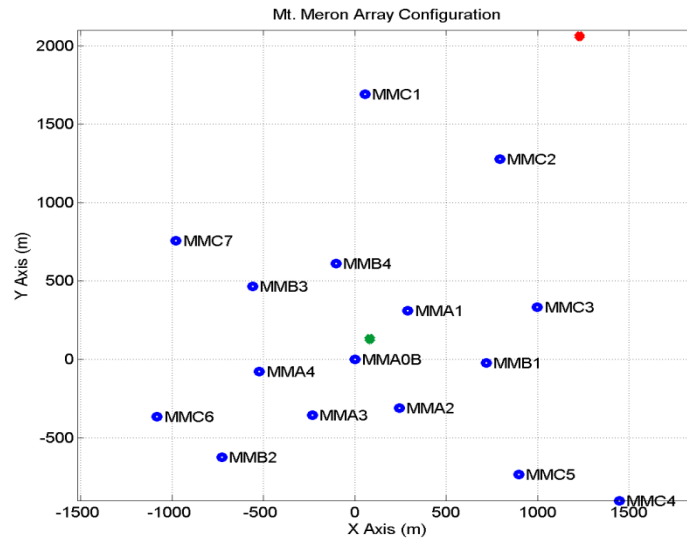


Figure 14

Sensor locations for Mt Meron array with reference location $(X, Y) = (80, 130) \text{ m}$ (green dot) and simulated source location $(X, Y) = (380, 640) \text{ m}$ (red dot).

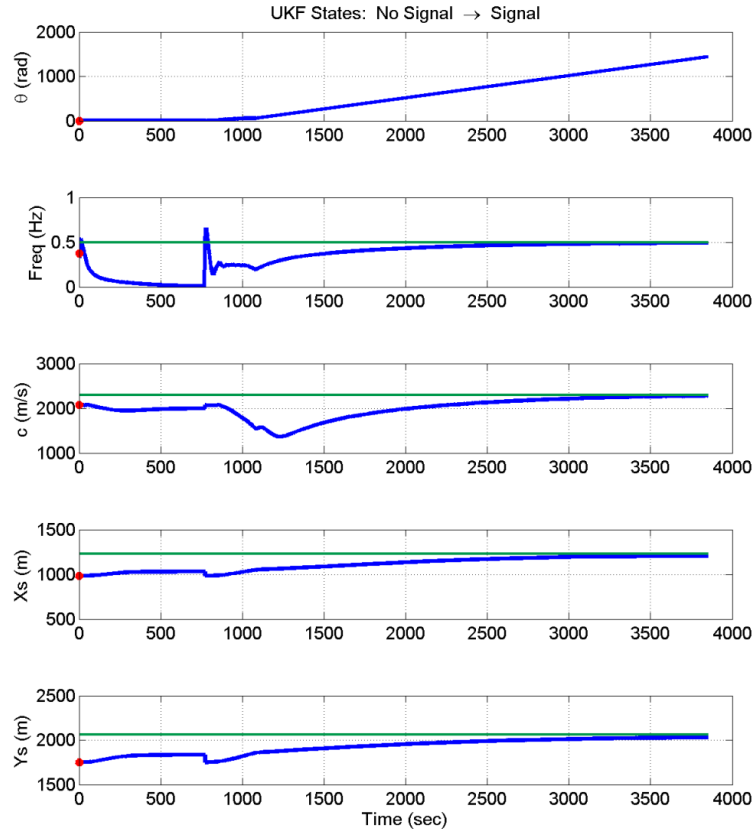


Figure 15

Time history of the states when the signal is turned on 766 seconds after the beginning of the record. The red dots at time zero are the initial state values. The green reference lines are the true values for each state.

Summary

We performed two types of initial analysis of the Mt Meron data for characterizing different noise components. In the first analysis, we estimated the local wave speed for frequencies less than 1.0 Hz by estimating the phase of the cross-spectrum for each data set. This showed wave speeds varying from 2100 m/s to around 7000 m/s. The noise in this band was propagating eastward, consistent with sources along the Mediterranean coast. An attempt to estimate speed and direction for noise above 1 Hz was not successful since the array spacing was too large to accurately resolve the phase differences between sensors.

In the second analysis we attempted to fit an unscented Kalman filter (UKF) to narrowband noise seen in the frequency band above 1 Hz. Testing of the narrowband noise model and UKF on simulated data using sources near the array showed promise. However, we did not obtain convergence of the filter on actual data. Two possible reasons are inadequate

spatial sampling (as mentioned previously), or lack of data preconditioning. This could be resolved by additional work if needed.

References

1. James V. Candy, *Bayesian Signal Processing*, IEEE Wiley, Hoboken, New Jersey (2009)